Group-wise penalized estimation schemes in model-based clustering

51st Scientific Meeting of the Italian Statistical Society

Alessandro Casa

Joint work with: A. Cappozzo & M. Fop

Faculty of Economics and Management
Free University of Bozen-Bolzano

alessandro.casa@unibz.it

23rd June 2022

Mixture modelling and clustering

- Model-based clustering offers a probabilistic formalization of the clustering problem
- Let $\mathbf{X} = \{x_1, \ldots, x_n\}$, with $x_i \in \mathbb{R}^p$, be the set of observed data. The density of a generic point is given by

$$f(x_i; \Psi) = \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k)$$

  - $\Psi = (\pi_1, \ldots, \pi_{K-1}, \theta_1, \ldots, \theta_K)$, with $\pi_k > 0$ and $\sum_k \pi_k = 1$
  - Gaussian component densities are often employed, hence $f_k(\cdot) = \phi_k(\cdot)$ with $\theta_k = \{\mu_k, \Sigma_k\}$
- MLE of Ψ is carried out via EM-algorithm and the partition is obtained resorting to the components-clusters correspondence

What about overparameterization?

- Major drawback: $|\Psi|$ scales quadratically with p, making this approach impractical in high-dimensional scenarios
- Several solutions to control the total number of parameters have been proposed:
  - Constrained modelling
  - Sparse estimation strategies
  - Variable selection
- We focus on the approach by Zhou et al. (2009), lying in between variable selection and sparse estimation methodologies

Penalized MBC

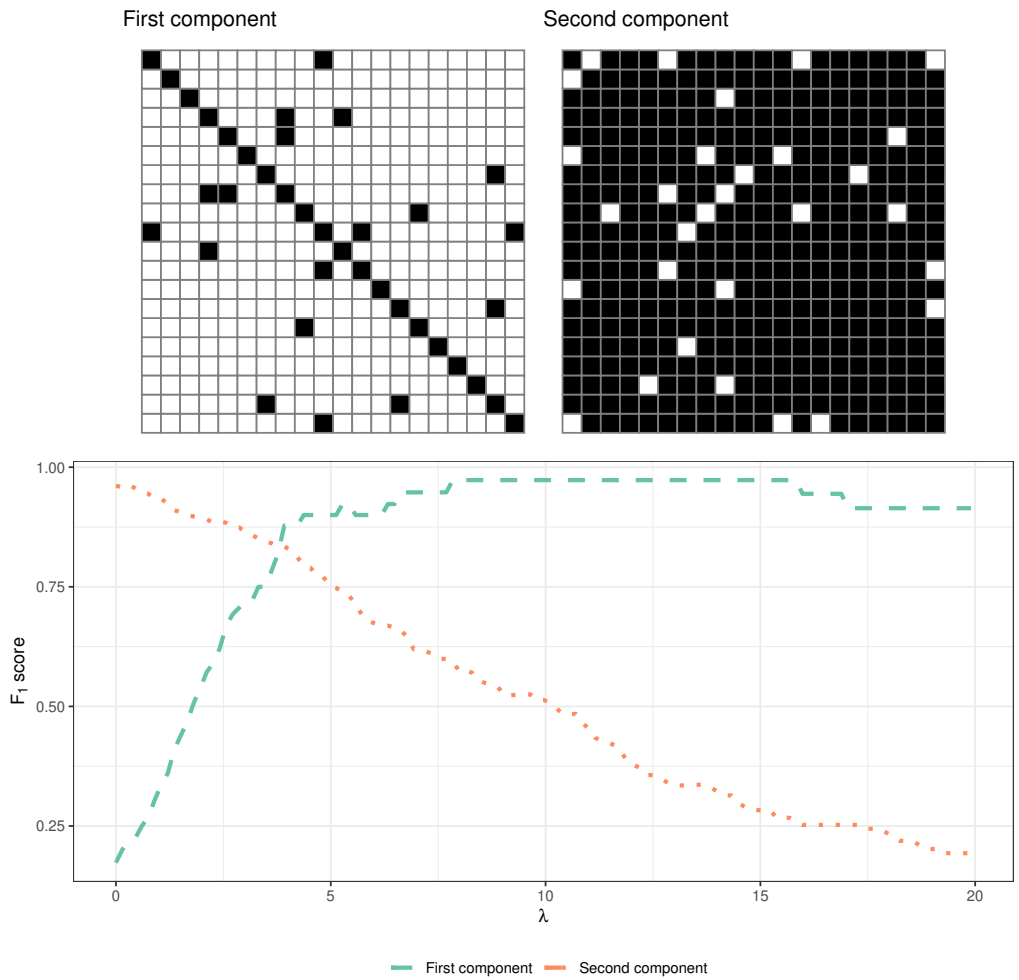
- Zhou et al. (2009) place a penalty on the component precision matrices $\Omega_k = \Sigma_k^{-1}$, to obtain sparser solutions
- Parameter estimates are obtained by maximizing

$$\tilde{\ell}_P(\Psi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Omega_k) - \lambda \sum_{k=1}^{K} ||\Omega_k||_1$$

  - The second term corresponds to the graphical lasso penalty applied class-wise, with $||\Omega_k||_1 = \sum_{jk} |\Omega_{jk}|$
  - A penalty on the mean vectors can be considered

A possible limit

- The mentioned approach implicitly assumes that classes have similar precision matrices structures
- What happens when we have under or over-connectivity?



Group-wise penalization in MBC

- We propose a method which overcomes the mentioned drawback by maximizing the following penalized log-likelihood

$$\ell_P(\Psi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Omega_k) - \lambda \sum_{k=1}^{K} ||\mathbf{P}_k * \Omega_k||_1$$

  with $*$ the Hadamard product and $\mathbf{P}_k$ weighting matrices

- Idea: we penalize transformations of the precision matrices, with $\mathbf{P}_k$'s encoding info about class-specific sparsity patterns

- Advantages:
  - Avoid the selection of K tuning parameters $\lambda_1, \ldots, \lambda_K$
  - Proper selection of $\mathbf{P}_k$ encompasses under or over-connectivity

And the weighting matrices?

- The interest is shifted toward the specification of $\mathbf{P}_1, \ldots, \mathbf{P}_K$
  Requirement $\longrightarrow$ stronger penalization on entries corresponding to weaker dependencies
- We define the matrices as

$$\mathbf{P}_k = f(\hat{\Omega}_k^{(0)})$$

  where $f : \mathbb{S}_+^p \longrightarrow \mathbb{S}^p$ and $\hat{\Omega}_k^{(0)}$ carefully initialized sample precision matrices
- Two approaches:
  - $P_{k,ij} = |\hat{\Omega}_{k,ij}^{(0)}|^{-1}$, inflate/deflate the penalty according to the entries $\hat{\Omega}_{k,ij}^{(0)}$
  - $P_k = \mathcal{D}\left(\hat{\Omega}_k^{(0)}, \text{diag}(\hat{\Omega}_k^{(0)})\right)^{-1}$, with $\mathcal{D}(\cdot, \cdot)$ an appropriate measure of distance

Model estimation

- For fixed K, λ and $\mathbf{P}_k$'s, the estimate $\hat{\Psi}$ is obtained maximizing $\ell_P(\Psi)$ by means of the EM-algorithm
- Sparse estimates for the component precision matrices are obtained by embedding graphical lasso in the M-step
- Model selection:
  - $\mathbf{P}_k$ is data-driven, does not require external tuning
  - Different values for K and λ are tested and the best combination selected maximizing

$$\text{BIC}_{\text{mod}} = 2 \log L(\hat{\Psi}) - d_0 \log(n)$$

with d_0 the number of parameters not shrunk to zero

Real data application

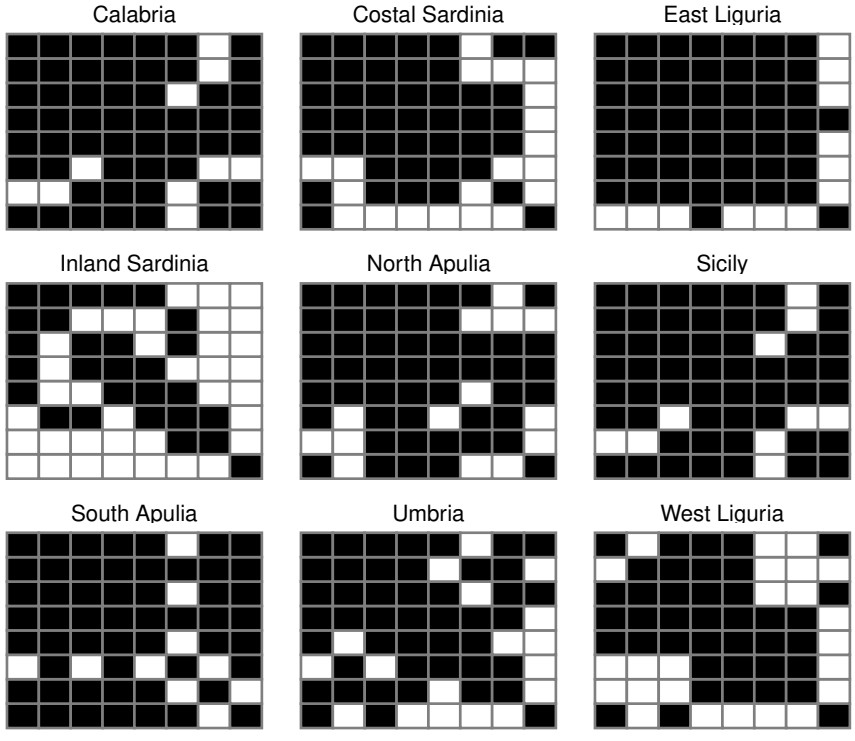
- Olive oil data
  $n = 572$ samples of olive oil, coming from $K = 9$ italian regions, with percentage composition measures of $p = 8$ fatty acids
- Aim: recover the group structure, given by geographical partition, using lipidic characteristics

- We compare our proposals with Zhou et al considering the following measures
  - ARI, checking if the clustering structure is recovered
  - d_Ω, the number of non-zero parameters in Ω_k's
  - MFD (Median Frobenius Distance), defined as

$$\underset{k \in 1,\ldots,K}{\text{median}} \left(||\hat{\Omega}_k - \bar{\Omega}_k||_F \right)$$

Some results - Olive oil

	ARI	d_Ω	MFD
Zhou et al.(2009)	0.6724	320	830
$\mathbf{P}_k$ via inverse $\lvert\hat{\Omega}_k^{(0)}\rvert$	0.7199	242	421
$\mathbf{P}_k$ via Frobenius dist	0.6875	312	701
$\mathbf{P}_k$ via Riemannian dist	0.6812	314	798



Some additional comments

- Additional numerical explorations, both on simulated and real data, produced further insights

- Good performances when no unbalancedness in Ω_k's sparsity
- The presence of the weighting matrices results in a procedure being less sensitive to λ selection
- Promising results (association and clustering structures recoveries) even when $p \geq n$

Conclusions and future work

- We generalize Zhou et al. (2009), by encompassing settings where the clusters have different amount of sparsity
- If paired with a penalty on the component means, the procedure can be used to perform variable selection

- And now?
  - Generalization to sparse covariance matrices estimation $\longrightarrow$ link with *Gaussian covariance graph model*
  - Extend the proposal to the Bayesian framework by borrowing concepts from *global-local shrinkage priors*

Some references

Casa, A., Cappozzo, A. & Fop, M. (2022+).
Group-wise shrinkage estimation in penalized model-based clustering.
Under review, `https://arxiv.org/pdf/2105.07935.pdf`

Other relevant references

- Bien, J. & Tibshirani, R.J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4), 807–820.
- Bouveyron, C. & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Comp Stat Data An*, 71, 52–78.
- Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.
- Zhou, H., Pan, W. & Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron J Stat*, 3, 1473–1496.